基于 AHP 和混合 Apriori-Genetic 算法的交通事故成因分析模型 *

邓晓衡，曾德天

(中南大学 信息科学与工程学院, 长沙 410075)

摘 要：针对交通事故数据多维多层的特点，对交通事故的主要成因与潜在规律进行了研究。从驾驶员、车辆、时间—地点、环境四个维度出发，提出了基于层次分析法（AHP）和混合 Apriori-Genetic 的模型挖掘事故成因。首先，引入 AHP 对事故诱发因素进行重要度排序，在客观分析的基础上将事故因素量化，筛选出引发交通事故的主要因素；其次，结合混合的 Apriori 和遗传算法对主要因素进行定向分析，找出关联规则，提高挖掘的准确性。相关对比实验的结果表明该模型可以减少无用规则的产生并提高挖掘的准确性，具有一定的科学意义和应用价值。

关键词：交通事故；层次分析法；Apriori；遗传算法；成因分析

中图分类号：TP391 doi: 10.3969/j.issn.1001-3695.2017.12.0818

Traffic accident causation analysis model based on AHP and hybrid Apriori-Gentic algorithm

Deng Xiaoheng, Zeng Detian

(*School of Information Science & Engineering*, *Central South University*, *Changsha* 410075, *China*)

Abstract: In view of the characteristic of multi-dimensional and multi-layer in traffic accident data, this paper proposed a new model to research the main reasons and potential rules in traffic accidents. The model starts from the four main dimensions such as the drivers, the vehicles, the time-address and the environment, and uses a way which based on AHP and hybrid Apriori-Gentic algorithm to mine causes of accident. First of all, the AHP sorted the importance of the influencing factors about accident. Then on the basis of objective analysis, the model quantified the influencing factors and selected the main influencing factors. Finally the model combined the genetic algorithm with the Apriori to directional analyze the main influencing factors and find the association rules out. The experimental result shows that the model could reduces the generation of useless rules and improves the accuracy of mining, which has certain scientific significance and application value.

Key words: traffic accident; AHP（analytic hierarchy process）; Apriori; genetic algorithm; causational analysis

0 引言

近年来，随着中国的汽车和驾驶人员数量高速增长，道路交通压力大增，交通事故有愈演愈烈的趋势[1]。同时中国是世界上交通事故死亡人数最多的国家之一，中国公安部官方最新的数据显示 2015 年全国交通事故发生总计 187781 起[2,3]。伴随着交通事故的产生，事故历史数据也逐步积累。为了探究交通事故的形成原因，利用数据挖掘技术对历史事故数据进行挖掘，希望能够找出数据中潜在的深层规则和数据模式，从而为交通事故的预防提供决策支持。由于在交通事故数据中诱发交通事故的因素众多，加上事故数据集中数据字段纷繁复杂且冗余信息很多，导致成因分析难以进行。为此本模型引入了层次分析法进行数据预处理。

层次分析法(AHP)[8]是运筹学家萨蒂提出的应用网络系统理论和多目标综合评价方法。它是一种将定性与定量相结合的系统分析方法，该方法将一个复杂问题分割成若干层，每层又包含若干因素，通过对事物的复杂本质和相关影响因素的深入分析后，绘制清晰的层次结构图，然后逐个的将各因素建立判断矩阵，通过计算判断矩阵的特征值和特征向量，得到不同因素的权重，根据权重值的大小评价结果，选出最佳的方案。

与此同时，为了提升挖掘的准确性，模型针对交通事故数据集将 Apriori 与遗传算法结合使用。Apriori 算法为布尔关联规则挖掘频繁项集算法[9]。它使用一种称作逐层搜索的迭代方法，k 项集用于探索 k+1 项集，直到不可能找到更大的频繁项集。通过频繁项集得出 A=>B 形式的关联规则，对于每一条规则主要有支持度和置信度两个参数进行衡量。

遗传算法通过模拟达尔文自然进化的思想来搜索全局最优解，它的初始种群是由随机产生的规则组成[10]。对每条染色体

收稿日期：2017-12-27；**修回日期**：2018-02-10 **基金项目**：国家自然科学基金资助项目（61379058，61772553）；湖南省科技计划项目（2015TP2017）；中南大学硕士生自主探索创新项目课题（2016zzts359）

作者简介：邓晓衡（1974-），男，湖南衡阳人，教授，博士，主要研究方向为数据挖掘、无线网络（dxh@csu.edu.cn）；曾德天（1993-），男，湖北仙桃人，硕士，主要研究方向为数据挖掘.

（规则）进行编码，由用户给出进化过程中的适应度函数，根据适者生存的原则，逐代优化，形成由当前群体中最适合的规则以及这些规则的后代组成的新群体。后代通过使用诸如交叉和变异等遗传操作来创建。

本文对城市交通事故成因分析展开研究，以贵阳市 2015 年全年交通事故历史数据[4]为分析依据，结合层次分析法和混合的 Apriori-Genetic 算法对数据进行建模分析，首先以 AHP 算法确定影响因素权重，选择主影响因子，剔除次要因素，简化运算；同时采用关联规则对主因素字段进行关联，通过遗传算法优化搜索结果。从而探究主因素背后的综合作用规律与交通安全的详细情况，如道路情况及其他交通环境对事故发生的影响，提高数据的利用价值。

1　相关工作

随着交通事故数据的大量积累，如何对数据建模研究并从中快速抽取出有用的信息成为了研究的重点工作。关联规则算法可以产生大量的关联规则，对于探究交通事故成因具有较好的适应性[5~7]。关联规则是使用支持度来寻找频繁项集，根据置信度发现关联规则，但在交通事故数据中由于字段过多，如果直接使用关联规则算法会产生大量无用且重复的频繁项与关联规则。

为了准确地找出交通事故的成因，需要对关联规则算法进行改进与优化，使其可以更好的应用到交通事故研究中来。遗传算法的搜索根据适应度函数进行，具有很强的方向性和目的性，可以弥补 Apriori 漫无目的的搜寻的缺陷。 Ghosh 等人[11]提出了基于遗传算法的频繁模式挖掘，他们将遗传算法引入到购物篮数据挖掘中，改良了挖掘的过程，在全局搜索的同时减少了时间复杂度，这一方法简单高效，同时在更大的数据集方面有更好的表现性。Chadokar 等人[12]提出了用于网络通信的混合关联规则与遗传算法，他们利用 Apriori 算法处理网络通信数据，得到频繁项集，之后再将频繁项集通过遗传算法得到更少更优的规则，实验通过对比单个 apriori 算法和混合算法的时间复杂度和产生的频繁项以及规则数量表明混合算法可以减少计算所花费的时间，并在产生的规则数量上更少，但规则质量更高。只是以上混合算法均针对特定场景设计，不具有普适性，如果要应用到交通事故成因分析中来，需要重新设计。

Jain 等人[13]提出了优化的关联规则挖掘算法，他们提出了正向的关联规则[14,15]挖掘和负向的关联规则[16,17]挖掘，对于每条规则引入了相关系数的概念，使用遗传算法来挖掘有效的正向关联规则与负向关联规则。这一方法可以减少搜索的空间同时通过每条关联规则所带的相关系数来判断此关联规则是否合适进一步的挖掘，但此模型在时间复杂度上效果并不理想。

2　模型设计

2.1　数据描述

本数据由贵阳市政府在 2016 年贵阳市交通大数据竞赛[4]中提供。原始的交通事故历史数据以 excel 文本的形式提供，共 56 651 条，含二十多个字段。事故成因种类分为 9 种，具体的分类见表 8。通过引入 2015 年贵阳市天气环境数据，形成了最终的数据集。由于数据中字段较多，无法直接开展关联规则分析。

2.2　层次分析法

层次分析法将定性定量相结合，迅速准确的找到问题的关键，并且能对各层次影响因素权重排序。故将其引入到本次成因分析的模型中来。

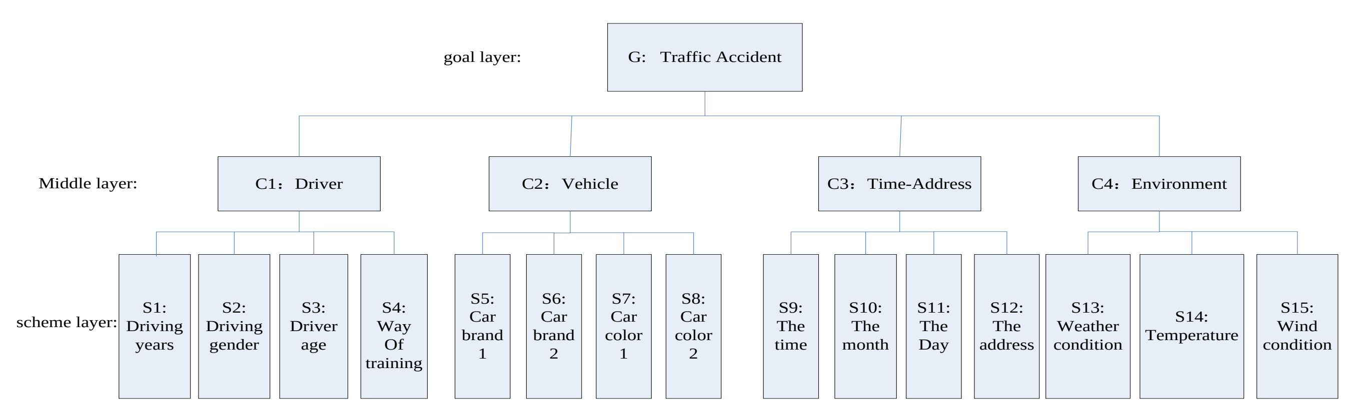


图 1　事故数据集中的系统层次结构

在交通事故数据集中，数据字段被大体分属为驾驶员类、车辆类、时间-地点类、环境类四个不同的维度，如图 1 中系统层次结构图所示。在与贵阳市通管理局及交通领域相关专家学者共同交流分析后，结合专家知识，参照 ahp 的 9 分位比率，构造了上下层次的判断矩阵，并进行矩阵一致性的检验；构造上下层次的判断矩阵见表 1~5。

表 1　目标(goal)层与中间(middle)层的判断矩阵 G-C 表

G	C1	C2	C3	C4
C1	1	5	3	2
C2	1/5	1	1/2	1/4
C3	1/3	2	1	1/2
C4	1/2	4	2	1

表 2　中间(middle)层与方案(scheme)层的判断矩阵 C1-S

C1	S1	S2	S3	S4
S1	1	5	3	2
S2	1/5	1	1/3	1/2
S3	1/3	3	1	2
S4	1/2	2	1/2	1

表 3　中间层(middle)与方案(scheme)层的判断矩阵 C2-S

C2	S5	S6	S7	S8
S5	1	1	1/2	1/3
S6	1	1	1/3	1/3
S7	2	3	1	1/2
S8	3	3	2	1

表 4　中间(middle)层与方案(scheme)层的判断矩阵 C3-S

C3	S9	S10	S11	S12
S9	1	3	2	4
S10	1/3	1	1/2	2
S11	1/2	2	1	3
S12	1/4	1/2	1/3	1

表 5　中间(middle)层与方案(scheme)层的判断矩阵 C4-S

C4	S13	S14	S15
S13	1	5	3
S14	1/5	1	1/3
S15	1/3	3	1

之后判断上述矩阵能否通过一致性检验，计算矩阵的最大特征值（如下所示）和其对应的特征向量，以及相应的一致性指标 CI 和检验系数 CR 值。

$$\lambda_{\max 1}=4.0211 \quad \lambda_{\max 2}=4.1074 \quad \lambda_{\max 3}=4.0458$$
$$\lambda_{\max 4}=4.0310 \quad \lambda_{\max 5}=3.0385$$

CI 代表矩阵的一致性，CI 越大，说明一致性越差；考虑到一致性的偏离可能是由于随机原因造成的，因此在检验判断矩阵是否具有满意的一致性时，还需将 CI 和平均随机一致性指标 RI 进行比较，得出最终检验系数 CR。计算公式分别如下所示：

$$CI=\frac{\lambda_{\max}-n}{n-1} \tag{1}$$

$$CR=\frac{CI}{RI} \tag{2}$$

平均随机一致性指标 RI 的值可以通过查平均随机一致性指标标准值表得到，它只和矩阵的阶数相关。当矩阵的阶数 n=3 时，RI 取 0.58；当矩阵的阶数 n=4 时，RI 取 0.90。对于每个矩阵，如果最终计算出来的 CR 值远小于 0.1，则说明矩阵的一致性检验通过，可以进行下一步的工作，否则重新分析构造矩阵。

表 6 列出了部分的计算结果及中间值，ω_k 代表相应矩阵最大特征值对应的特征向量。最终可以得出方案层中每个属性相对于目标层的权重值，选取权重大于某一阈值的字段作为影响交通事故的主要因素，现经过经验测试选取合适的阈值为 0.044。

表 6　部分计算的结果及中间值

	G-C	C1-S	C2-S	C3-S
ω_{k1}	0.4773	0.4909	0.1377	0.4673
ω_{k2}	0.0809	0.0863	0.1258	0.1601
ω_{k3}	0.1539	0.2483	0.2879	0.2772
ω_{k4}	0.2880	0.1745	0.4486	0.0954
$\lambda_{\max}$	4.0211	4.1074	4.0458	4.0310
CI	0.007	0.0358	0.0153	0.0103
CR	0.0078	0.039	0.017	0.0011

通过层次分析法对照表 7 中的数据可以发现准则层中的 Driver 相对目标层 Traffic Accident 所占权重为 0.4773，而方案层中 Driver age 相对 Driver 所占权重为 0.4909，故 Driver age 字段相对于 Traffic Accident 所占的权重为 0.4773*0.4909=0.2343；同理将方案层中每个字段相对于目标层 Traffic Accident 所占权重依次计算，选取最终权重大于阈值的字段做为主要事故影响因素。

表 7　各字段属性相对目标层的权重值排列

criterion layer	Weight	scheme layer	Weight
		Driving years	0.4909
Driver	0.4773	Driving gender	0.0863
		Driver age	0.2483
		Way of training	0.1745
		Car brand 1	0.1377
Vehicle	0.0809	Car brand 2	0.1258
		Car color 1	0.2879
		Car color 2	0.4486
		The day	0.2772
Time-	0.1539	The month	0.1601
Address		The time	0.4673
		The address	0.0954
		Weather condition	0.6369
Environment	0.2880	Temperature	0.1047
		Wind condition	0.2583

2.3　混合的 Apriori-Genetic 算法

结合遗传算法的优点，本文设计了一种针对交通事故成因分析的混合遗传关联规则挖掘算法，采用 Apriori 来发现输入数据中的频繁项集。通过将频繁项按某种形式进行编码转换为染色体，将这批染色体作为遗传算法的初始种群，再根据预定义的适应度函数对每条染色体计算其适应值，通过选择适应值高的一批染色体进行复制，通过遗传操作（选择，交叉，变异）

产生新的一代群体。通过不断的繁殖进化，最后收敛到一批具有较高适应度的个体上或者迭代的次数达到了预设定的阈值，即输出最优分类规则集。混合算法的流程图如图 2 所示。

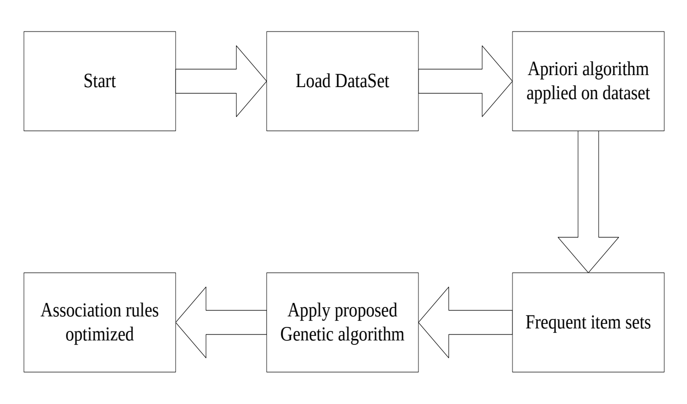


图 2　混合算法的总体流程图

2.3.1 编码设计

对于交通事故数据集，将影响交通事故的因素作为规则前件部分，将交通事故原因类型作为规则后件。期望能够找到"驾龄、年龄、培训驾校、时间等字段=>事故原因类型"这一形式的规则，规则的前件中每一个特征属性（如驾龄）有 n 个分类，则使用相应 x 位二进制进行表示，其中 x 与 n 满足关系：

$$\min\{x \mid 2^x > n\} \tag{3}$$

规则的后件"事故原因类型"作为分类属性，代表造成事故的原因，事故原因共有 9 种（见表 8），用二进制表示的方法同规则前件中的特征属性。表 9 描述了数据中驾龄字段的分类。

表 8　事故原因类型分类与对照表

Label	Mean
1	追尾的
2	逆行的
3	倒车的
4	停车时未挂低速档、未拉驻车制动，导致车辆滑行的
5	开关车门的
6	违反交通信号的
7	未按规定让行的
8	依法应负全责的其他情形
9	不符合前 8 款规定或者双方同时具有上述情形的

通过 Apriori 算法得出事故频繁项集，挑选其中同时含有特征属性与分类属性的项集，作为初始的规则进行编码。例如频繁项：[培训方式 ='驾校培训', 驾龄 ='驾龄 1', 事故原因类型 ='1']，其中培训方式有’驾校培训’与’自培’两类，则式（5）中 n=2，x 最小取 2；可设'驾校培训'对应编码为’01’，’自培’对应’10’，若此属性未出现在此频繁项中则对应的二进制为‘00’，其他字段的编码设计同理。在程序中构造了一个列表用于存放规则所对应的二进制染色体。列表的长度为 23，对应 AHP 方法筛选出的 7 个特征属性和 1 个分类属性，每个属性字段所对应的编码按固定顺序依次存于列表当中。

表 9　驾龄字段的分类与对照表

Label	Mean
驾龄 1	0-4 years
驾龄 2	5-11 years
驾龄 3	12-19 years
驾龄 4	20 years and more

2.3.2 定义适应度函数

适应度函数是用来评价个体适应环境的能力，是进行自然选择的依据。对于期望的规则可以使用支持度，置信度，覆盖度等多种指标进行评价。在适应度函数设计中，基于综合考虑，令适应度函数

$$F(r) = a*S(r) + b*C(r) + c*CR(r) \tag{4}$$

其中：变量 r 代表规则，a，b，c 均为常量系数并且 a，b，c 的取值范围为[0,1]。令 N 为整个数据集的记录数，C 为规则中除去‘事故原因类型’属性后的其他字段，C 在 N 中出现的频数用 R_C 表示；D 表示‘事故原因类型’字段，D 在 N 中出现的频数用 R_D 表示；C，D 同时出现在数据集中的频数计为 $R_C \cup R_D$，S(r)为规则的支持度，则 S(r)的定义为

$$S(r) = \frac{R_C \cup R_D}{N} \tag{5}$$

C(r)为规则的置信度，C(r)的定义为

$$C(r) = \frac{R_C \cup R_D}{R_C} \tag{6}$$

同理规则的覆盖度 CR(r)定义为

$$CR(r) = \frac{R_C \cup R_D}{R_D} \tag{7}$$

常量系数 a，b，c 是本模型的关键所在，可由用户根据需要进行调整，从而对规则评价的偏重可以发生相应的改变，使得进化沿用户期望的方向进行，提高挖掘的准确性。

2.3.3 遗传算子设计

1)选择算子

选择操作使用轮盘赌操作，其具体过程描述如下所示：对于 Apriori 算法选出的初始种群中的每个染色体，计算其适应度值，将所有的适应度值刻画到一个圆盘上，即适应度值的大小表示在圆盘上的面积。在转动轮盘的过程中，单个染色体的面积越大，则被选中的概率越大。

2)交叉算子

从 ICGA 和 GP 会议的历年相关文献来看，交叉概率的选取并无固定的方法与逻辑，一般取 0.4~0.99；但交叉概率取值过大，则不利于种群中优秀基因的保存，取值过小就会导致种群进化缓慢。在本次交通事故数据处理中经反复测试，交叉概率为 0.6 时，对于进化速度和实验结果能有一个较好的表现。

故设置交叉概率为 0.6，为了在不破坏种群的基因多样性的前提下加快种群的进化速度，使用选择算子选择出父本和母本后，按单点交叉随机产生交叉位，形成两个新的个体，考虑到

在挖掘中为了找到更优的规则，将新产生的个体按适应度排序，再从中挑选出大于适应度阈值的个体加入到解中；同时也将这些挑选出的个体加入到原先的种群，从而丰富种群。这样既保存了父本和母本的基因，又在进化的过程中保持了种群的多样性。

3)变异算子

在遗传算法中使用可变的变异概率，设 P 为变异概率。具体描述如下：

if (个体的适应度>群体的平均适应度)

then {P 取一个相对较小的值或接近 0;}

else {P 值取一个相对较大的值;}

3　实验结果

3.1　实验设计

采用保留“事故原因类型”属性类别比例的分层采样，随机采样选取数据集中 70%的数据作为测试集，用来寻找关联规则；30%为验证集，验证生成的关联规则在验证集上的置信度。层次分析法最终选取出了驾龄、性别、年龄、培训驾校、时间、天气状况、风力风向共 7 个字段作为事故成因分析的主要影响因素，和事故原因类型字段一起作为混合的 Apriori 和遗传算法的输入。分别使用单独的 Apriori 算法，单独的遗传算法以及混合的 Apriori-Genetic 算法对事故历史数据进行处理，比较数据挖掘结果并进行性能测试。同时在找到的期望规则数量上，将混合算法与 C4.5 及随机森林等其他的机器学习算法进行比较，以验证其性能。实验环境为 intel Core i5-4200H 处理器，8 GB 内存，Windows7 操作系统，python 语言。

3.2　混合算法的挖掘结果

为便于比较，令适应度函数 F(r)（参见式（4））中常系数 a=b=c=1。对于预处理后的数据集实施混合的 Apriori-Genetic 算法，支持度阈值设为 0.1，在表 10 中列出了使用本文提出的模型找到的部分适应度较高的期望规则，规则后面分别附有相应的适应度，支持度，置信度，和覆盖度。

第一条规则'男性','8-12 点'=>'事故类型 1'的适应度为 1.72，其支持度为 0.28，置信度为 0.71，覆盖度为 0.73，从规则中可以看出驾驶员为男性&时间为上午 8-12 点的组合中，经常会出现事故原因为追尾的事故，且规则具有较高的覆盖度。规则'0-4 年驾龄','女性'=>'事故类型 4'的适应度为 1.71，它显示了女性驾驶员在驾龄偏低的情况下，易发生类型 4 的事故，这也显示了年轻女性司机在技术上还需多加练习。规则'自培'，'雨天'=>'事故类型 7'显示，对于驾驶员为自培形式的在雨天容易发生未按规定让行的事故，说明未经驾校培训的驾驶员在交通规则的学习上需要加强。而当驾驶员为驾校培训&18-25 岁的男性时，出现事故原因为未按规定让行的事故概率亦较高。推测可能是年轻人驾驶技术不娴熟导致。通过详细分析得出的交通事故规则结果集，可以找出有意义的组合规则，对于交通事故的针对性预防和科学的管理具有重要的意义。

当用户对规则评价的偏重发生了改变，如更加关注置信度时，则可以使置信度的系数 b 相对 a 和 c 而言较大，最终得出的期望规则集会在置信度上有更好的表现。此处以 a=1>b=c=0.5；b=1>a=c=0.5 以及 c=1>a=b=0.5 进行三组不同偏重下的规则挖掘实验，得出的适应度最高的期望规则集分别如表 11~13 所示。

表 10　a=b=c=1 条件下挖掘规则结果表

规则	适应度	支持度	置信度	覆盖度
'male','time2'=>'1'	1.72	0.28	0.71	0.73
'Driving experience 1', 'female'=>'4'	1.71	0.32	0.88	0.51
'school training '=>'1'	1.66	0.37	0.79	0.54
'male', 'rain'=>'1'	1.73	0.44	0.67	0.62
'self training', 'rain'=>'7'	1.91	0.41	0.83	0.67
'self training','male', 'time 3'=>'7'	1.55	0.33	0.75	0.47
'school training ', 'male', 'age 2'=>'7'	1.32	0.25	0.71	0.36

表 11　a=1，b=c=0.5 条件下挖掘规则结果表

规则	适应度	支持度	置信度/2	覆盖度/2
'male','time2'=>'1'	1.00	0.28	0.355	0.365
'Driving experience 1', 'female'=>'4'	1.015	0.32	0.44	0.255
'school training '=>'1'	1.035	0.37	0.395	0.27
'male', 'rain'=>'1'	1.085	0.44	0.335	0.31
'self training', 'rain'=>'7'	1.16	0.41	0.415	0.335
'self training','male', 'time 3'=>'7'	0.94	0.33	0.375	0.235
'school training', 'Northeasterly wind', 'rain'=>'6'	**0.881**	**0.42**	**0.298**	**0.163**

表 12　b=1，a=c=0.5 条件下挖掘规则结果表

规则	适应度	支持度/2	置信度	覆盖度/2
'male','time2'=>'1'	1.212	0.14	0.71	0.365
'Driving experience 1', 'female'=>'4'	1.295	0.16	0.88	0.255
'school training '=>'1'	1.245	0.185	0.79	0.27
'male', 'rain'=>'1'	1.2	0.22	0.67	0.31
'self training', 'rain'=>'7'	1.37	0.205	0.83	0.335
'Driving experience 1', 'rain'=>'3'	**1.173**	**0.097**	**0.79**	**0.286**
'self training','male', 'time 3'=>'7'	1.15	0.165	0.75	0.235
'school training ', 'male', 'age 2'=>'7'	1.015	0.125	0.71	0.18
'Driving experience 2','time 3'=>'7'	**1.037**	**0.112**	**0.82**	**0.105**

将表 11 与 10 对比分析可知，表 11 结果集中新增了规则（黑体加粗部分）：'驾校培训'，'东北风'，'雨天' => '事故类型 6'；由天气历史数据可知，贵阳地区的风向长期以冬季的东北风和夏季的西南风为主，针对此规则建议可以加强驾校培训学员在冬季起东北风且雨天时的安全教育；此规则的支持度较高，置信度为 0.596，但覆盖度的值偏低，其余规则与表 10 基本保持

一致。

将表 12 与 10 对比分析易知，在重点关注置信度的情况下，结果集中同样找到了两条适应度较高的新规则：'0-4 年驾龄'，'雨天'=>'事故类型 3'以及'5-11 年驾龄'，'下午 13-18 点'=>'事故类型 7'，它们都具有较高的置信度，但由于其他指标值不高的原因，导致在系数全部相等时，在混合算法的挖掘中适应度不高而未展现出来，其余规则亦与表 10 基本保持一致。将表 13 与 10 对比分析可知，在重点关注覆盖度的情况下，发现了一条适应度较高的新规则：'女性'，'自培'，'雨天'=>'事故类型 1'，其余规则同表 10 保持一致。

对于适应度函数中系数的不同设置体现了用户对于不同指标的关注程度，而用户的偏重会在接下来混合算法的挖掘结果中有所体现，即可能找到一些新的期望规则，从而达到定向挖掘的目的，提高模型分析的精确性。

表 13　c=1，a=b=0.5 条件下挖掘规则结果表

规则	适应度	支持度/2	置信度/2	覆盖度
'male','time2'=>'1'	1.225	0.14	0.355	0.73
'Driving experience 1', 'female'=>'4'	1.11	0.16	0.44	0.51
'school training '=>'1'	1.12	0.185	0.395	0.54
'male', 'rain'=>'1'	1.175	0.22	0.335	0.62
'self training', 'rain'=>'7'	1.29	0.205	0.415	0.67
'self training','male', 'time 3'=>'7'	1.01	0.165	0.375	0.47
'female', 'self training', 'rain'=>'1'	**1.059**	**0.089**	**0.31**	**0.66**

3.3　与其他算法的性能比较

对于混合算法找出的规则在验证集中计算其相应的适应度值，图 3 显示了规则在测试集与验证集中适应度的比较。可以看出，除规则 r6 差别较大外，其他规则总体较为接近，说明上述找到的规则是可靠的。

接下来只使用 Apriori 算法对 AHP 选出的字段数据进行挖掘，得出关联规则；只使用遗传算法对选出的字段数据进行挖掘，初始种群随机生成，其种群大小和混合算法的初始种群相等，函数适应度的设计同式(4)，对于遗传算法中的遗传算子的设计与 Apriori-Genetic 混合算法保持一致。通过设置不同的支持度和遗传代数对比三种算法的实验效果。

图 4 上图所示是混合算法与单独 Apriori 算法在不同支持度下产生的期望规则(这里定义“期望规则”的适应度值大于 1.0，从而保证规则的可靠性)在数量上的比较，其中混合算法的遗传代数为 100；可以看出在相同的支持度下，由于混合算法使用了支持度、置信度、覆盖度作为规则的评价指标，加上遗传算法的优化，故混合算法能找到更多且更符合用户期望的规则，而 Apriori 在支持度增大时，得到的期望规则却逐渐减少。图 4 下图所示是混合算法与单独遗传算法在相同的支持度（0.1）不同的遗传代数下产生的期望规则数量比较。而在遗传代数较少时，由于单独遗传算法的初始种群为随机生成，发现的期望规则数相对混合算法较少，但随着遗传代数的增加，搜索空间增大，简单遗传算法和混合算法所找到的规则数趋于接近，效果相差不大。

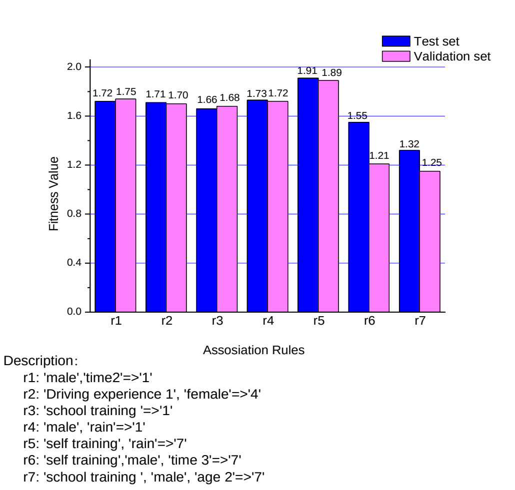
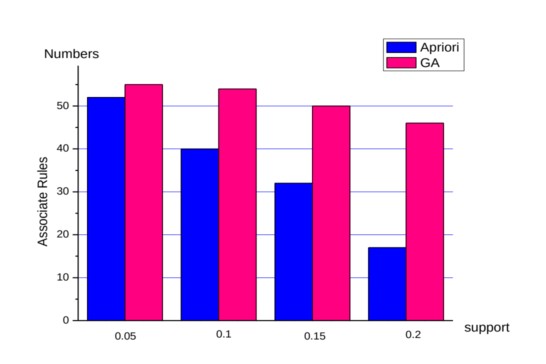


图 3　相关期望规则在测试集与验证集上适应度的比较



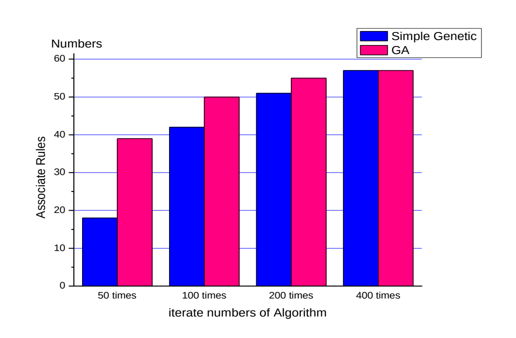


图 4　不同算法下生成期望规则在数量上的比较

同时对比混合算法与 C4.5，随机森林等其他算法获得的期望规则数量，图 5 显示了在不同支持度阈值下不同算法获得的期望规则数量。数据显示混合算法的寻找能力要优于 C4.5 与随机森林算法，验证了混合算法的优秀性能。

图 6 所示是混合算法与单独 apriori 算法在不同支持度下运行结束的时间比较，以及混合算法与单独遗传算法在相同的支持度（0.1）不同的遗传代数下运行结束的时间比较。由图可知当遗传代数较少时混合算法的表现要优于单独的遗传算法；但混合算法的运行时间要差于单独的 Apriori 算法。

为了降低混合算法的时间复杂度，在读取数据并计算适应度函数时通过开启多线程协作，可以降低算法的时间复杂度。

图 7 所示是并行化后的混合算法与单独 apriori 算法在不同支持度下运行的时间比较，以及并行化后的混合算法与单独遗传算法在相同的支持度（0.1）不同的遗传代数下运行的时间比较。由图可知混合算法在并行化后的运行时间有较大改善，接近于单独的 Apriori 算法。

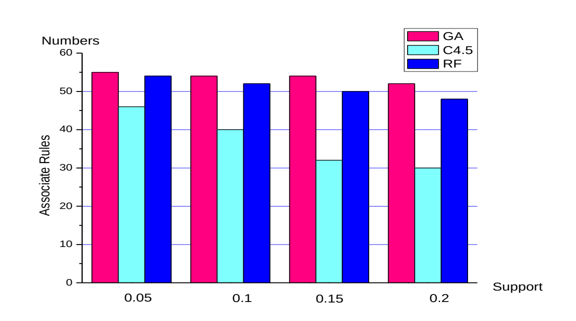


图 5　不同算法下获得的期望规则在数量上的比较

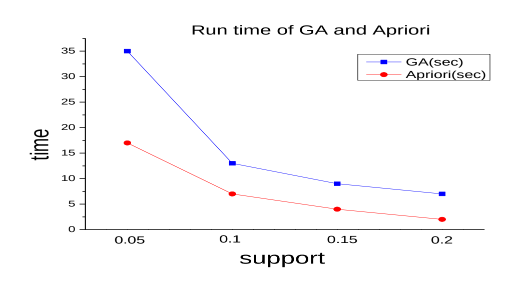
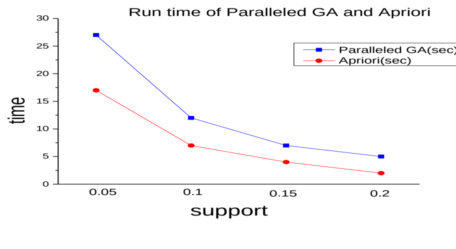


图 6　不同算法在运行时间上的比较（单位：s）



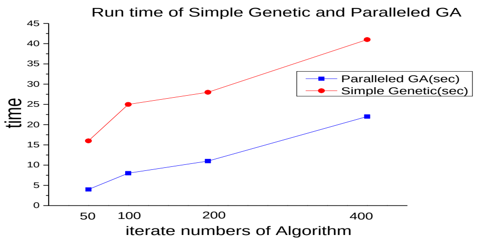


图 7　混合算法并行化后与其他算法在运行时间上的比较（单位：s）

4　结束语

本文主要面向交通事故数据集的数据挖掘工作，针对交通事故数据复杂的多维多层特点，在对比现有事故处理算法的基础上，引入了 AHP 方法，同时设计了混合的 Apriori-Genetic 算法用于挖掘交通事故成因，并建立了事故成因分析模型，挖掘事故成因。通过对比混合算法与传统算法的性能，表明混合算法具有可行性，同时实验结果表明该模型可以提高挖掘的准确性并减少无用规则的产生，具有较好的应用价值。但 AHP 矩阵的构造终究带有一定的主观性，如何消除这一误差将是下一步的研究工作。

参考文献：

[1] 中华人民共和国交通运输部. 2014 中国交通运输统计年鉴 [M]. 北京: 人民交通出版社, 2015: 11-15.

[2] 中华人民共和国公安部. 交通事故发生数总 (起) [EB/OL]. (2015-6-17) [2017-7-8]. http://data. stats. gov. cn/search. htm?s=trafficaccident

[3] 中华人民共和国公安部. 中华人民共和国道路交通事故统计报告 [EB/OL]. (2016-8-5) [2017-7-9]. http://www. mps. gov. cn/n2256342/index. html.

[4] 贵阳市政府. 贵阳交通大数据竞赛 [EB/OL]. (2015-7-23) [2017-7-12]. http://jjzd. gygov. gov. cn/art/2016/1/15/art_24609878350. html.

[5] Tibebe B. Mining road traffic accident data to improve safety: role of road-related factors on accident severity in ethiopia [C]// Proc of AAAI Spring Symposium on Artificial Intelligence for Development. 2010: 377-384.

[6] Marukatat R. Structure-based rule selection framework for association rule mining of traffic accident data [C]// Proc of International Conference on Computational and Information Science. Berlin: Springer, 2006: 231-239.

[7] Murat Y S. Modelling traffic accident data by cluster analysis approach [J]. TEKNIK DERGI, 2009, 20 (3): 759-777.

[8] Roshamida A J, Amir M A U. Risk assessment of dry bulk cargo operations using Analytic Hierarchy Process (AHP) method [J]. IEEE Trans on Information and Communication Technology, 2016: 146-159.

[9] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases [J]. ACM SIGMOD Record, 1993: 207-216.

[10] Whitley L D. The GENITOR algorithm and selection pressure: why rank-based allocation of reproductive trials is best [C]// Proc of the3rd International Conference on Genetic Algorithms. San Francisco: Morgan Kaufmann Publishers Inc, 1989: 116-123.

[11] Ghosh S, Biswas S, Sarkar D, *et al.* Mining frequentitemsets using genetic algorithm [J]. International Journal of Artificial Intelligence & Applications, 2010, 1 (4) .

[12] Chadokar S K, Singh D. Optimizing network traffic by generating association rules using Hybrid Apriori-Genetic algorithm [J]. IEEE Trans. on Wireless and Optical Communications Networks, 2013: 1-5.

[13] S. Jain. Mining & optimization of association rules using effective algorithm [J]. International Journal of Emerging Technology and Advanced Engineering, 2012, 2 (4): 281-285.

[14] Kishor P, Porika S. An efficient approach for mining positive and negative association rules from large transactional databases [J]. IEEE Trans on Inventive Computation Technologies, 2016, 1: 74-77.

[15] Naredi S, Deshmukh R A. Improved extraction of quantitative rules using Best M Positive Negative Association Rules Algorithm [J]. IEEE Trans on Electronics, Computing and Communication Technologies, 2015: 13-18.

[16] Ravi C, Khare N. EO-ARM: an efficient and optimized k-map based positive-negative association rule mining technique [J]. IEEE Trans on Circuit, Power and Computing Technologie, 2014: 1723-1727.

[17] Doshi M, Roy B. Enhanced data processing using positive negative association mining on AJAX data. Circuits [J]. IEEE Trans on Systems, Communication and Information Technology Applications, 2014: 386-383.